The Contribution of Nasal Murmur to the Perception of Nasal Consonant

Fan Bai[1,2,*]

[1]Department of Electronic Engineering, University of Electronic Science and Technology of China, Chengdu, Sichuan, 611731, P.R. China

[2]Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Urbana, Illinois, 61801, USA

Abstract: Identification of perceptual cues can be very helpful in almost all areas of speech signal processing. Recently, a new methodology called the 3-Dimensional-Deep Search and a visualized intelligible time-frequency computer-based model AI-gram have been introduced for research on the perceptual cues. Based on the technique, the acoustic cues for stop consonants [1], fricative consonants [2] and nasal consonants [3] are successfully found. However, these have limitations for studying the contribution of nasal murmur to the recognition of nasal consonants due to the following reasons: Firstly, they only allow the investigation of individual recognition effects along the time, frequency and amplitude axes. The effects of frequency and amplitude in a combinatorial way cannot be studied. Secondly, the initial value for the high-pass filter in the filter experiment HL07 [4] is set to 697 Hz, but the nasal murmur region lies around 250 Hz. The perceptual contribution of nasal murmur to the nasal consonants cannot be assessed. To solve these problems, a new experiment has been designed by analyzing the experiment data and comparing them with the stimuli under different SNRs via AI-gram. It is revealed that when the primary cue of nasal consonant is clear, which is usually under high SNRs, filtering out nasal murmur does not affect its correct perception. However, when the primary cue is weak usually under low SNRs, nasal murmur has strong complementary effects on the primary cue, and can greatly suppress confusions. This conclusion can be used for noise-robust speech recognition.

Keywords: Nasal murmur, perceptual cue, nasal consonant, AI-gram.

1. INTRODUCTION

Acoustic cues or perceptual cues are time-varying spectral patterns which characterize the perception of every single syllable, and are used by the auditory system to decode the phoneme [5, 6]. Perceptual cues can be used in almost all branches of speech signal processing, such as speech compression, enhancement, automatic speech recognition (ASR) and hearing aid, thus it is quite meaningful to find stable cues. The first research on acoustic cues was performed as the visible speech project in Bell Labs (1940) by Potter *et al.* [7], where they trained the hearing-impaired to read spectrograms. Following this, a lot of research [8-14] has been conducted to search for acoustic cues of consonants. To control the variability of speech cues in perceptual experiments, almost all of these researches on feature analysis were carried out using synthetic speech. Unfortunately, speech synthesis relies on prior knowledge or assumptions about the speech cues, however, this knowledge or assumptions may not be accurate and sufficient enough to produce natural and highly intelligible speech sounds [15, 16]. On the other hand, the variability of natural speech makes it very hard to study perceptual cues. The same speech produced by different talkers of different accents or gender may have different perceptual cues [17].

To overcome these problems and to obtain real human speech acoustic cues, the Human Speech Recognition (HSR) research group at the University of Illinois Urbana-Champaign has developed a methodology called the 3-Dimensional-Deep Search (3DDS) [1] with human nature speech as stimuli.

1.1. The Basic Principle of 3DDS and AI-gram

The basic principle of 3DDS methodology is to systematically remove various parts of a speech sound along three axes and then to assess the significance of the removed component from the change in the recognition score along three dimensions: time, frequency, and signal-noise ratio (SNR) (see Fig. **1**).

Based on the results of three corresponding psychoacoustic experiments, 3DDS method helps to allocate acoustic cues in AI-gram [18-20], which is a new visualized intelligible time-frequency computer-based model. Given a speech sound and masking noise, the AI-gram simulates the effect of noise masking and produces an image that predicts the audible speech components along the time and frequency axes. The block diagram of how an AI-gram is computed given a noisy speech signal s(t) is shown in Fig. (**2**).

*Address correspondence to this author at Department of Electronic Engineering, University of Electronic Science and Technology of China, Chengdu, Sichuan, 611731, P.R. China; Tel: +001 2174189782;
E-mail: baifan11111@gmail.com

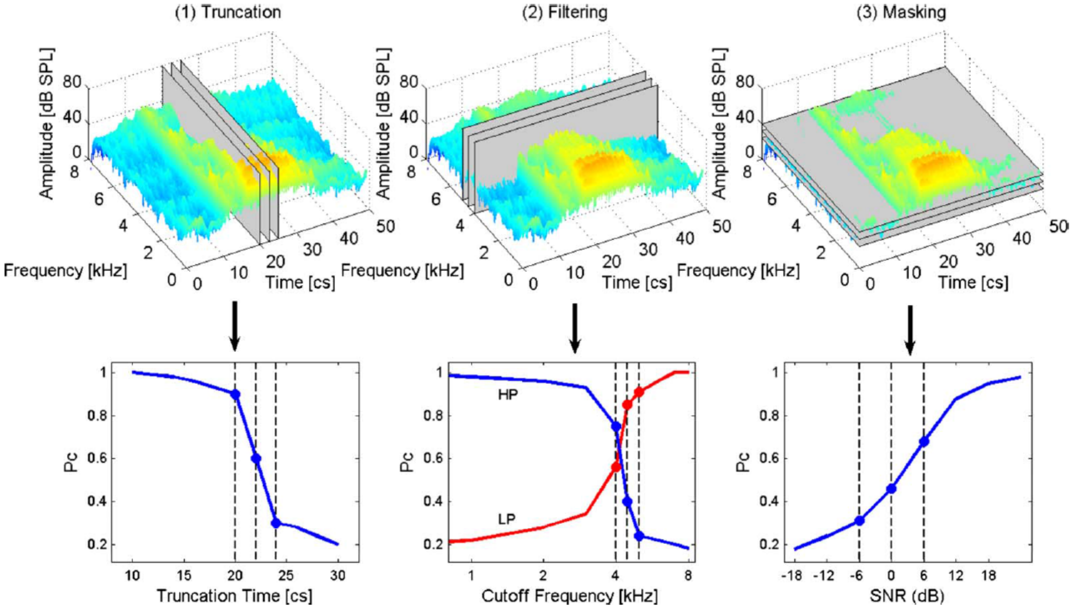


Fig. (1). The 3DDS for the identification of acoustic cues [1].

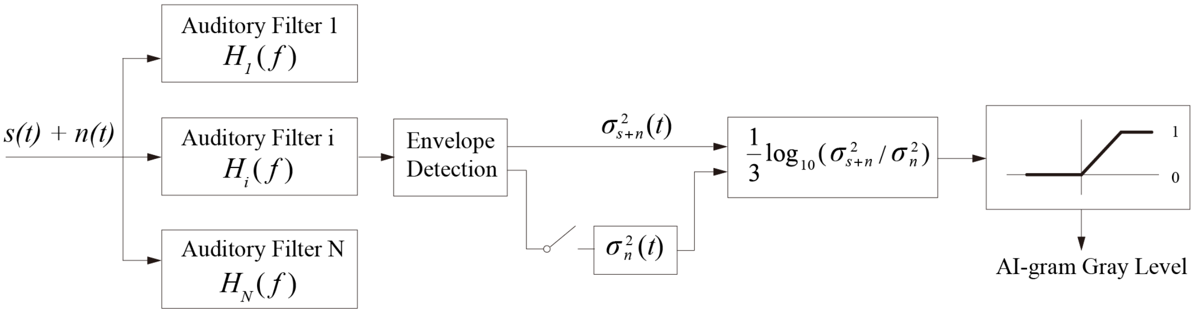


Fig. (2). Block diagram of AI-gram (modified from Li *et al.*, 2010).

The 3DDS methodology has already been successfully used for locating the acoustic cues of stop consonants [1], fricative consonants [2] and nasal consonants [3].

1.2. Previous Study in the Cue of Nasal Consonants by 3DDS

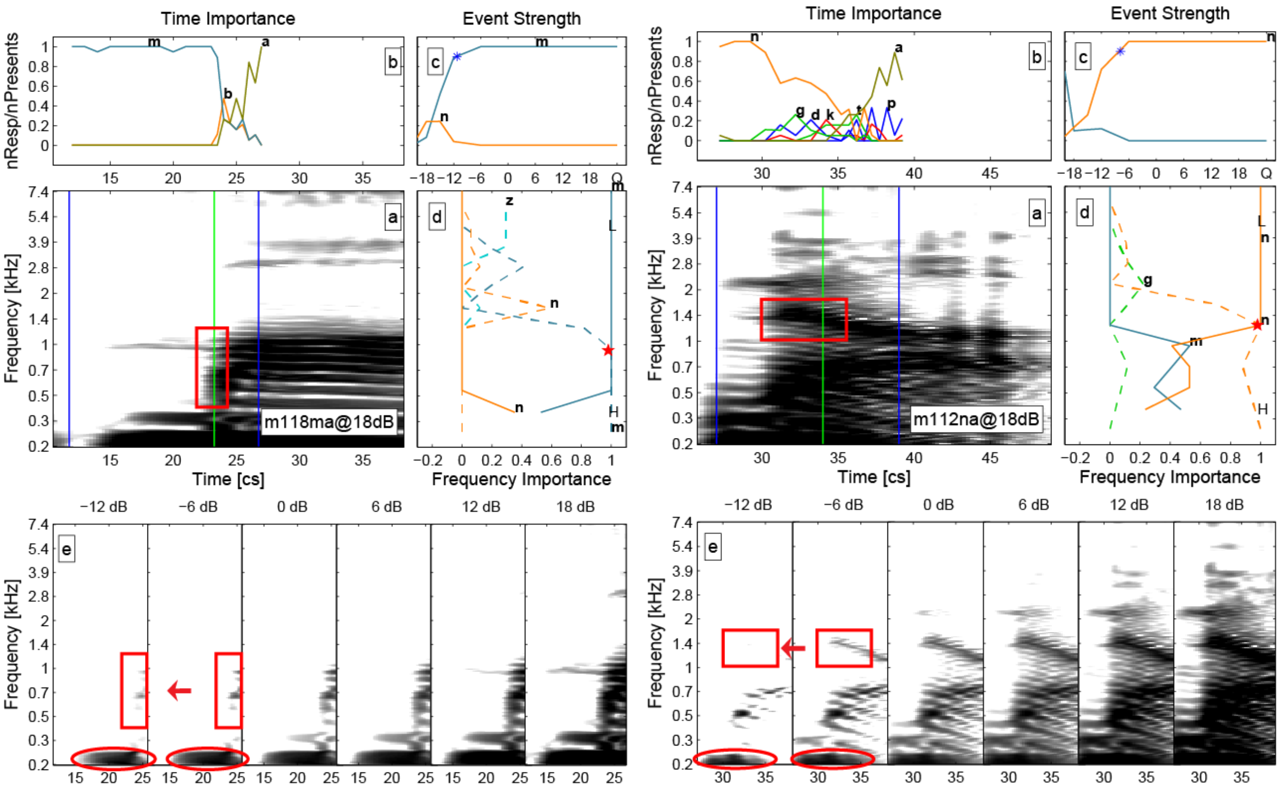
The primary cues of nasal consonants /m/ and /n/ by 3DDS [3] have been found. The perceptual cue of /ma/ is the onset of the F2 formant region ranging between 0.35 and 1.2 kHz as highlighted in red rectangles in Fig. (**3A**). The perceptual cue of /na/ is an F2 transition region around 1.5 kHz as highlighted in red rectangles in Fig. (**3B**). Since the first cut-off frequency of high-pass filter in the high/low pass filter experiment [4] is 697 Hz, the perceptual contribution of nasal murmur cannot be assessed, which lies around 250 Hz. Also, because the three psychoacoustic experiments on time, frequency and SNR are independent [1], the perceptual effect of removing nasal murmur under different noise conditions cannot be evaluated. It is also meaningful to evaluate the effect cue for such manner. To solve this problem and investigate the perceptual contribution of nasal murmur, an experiment had been performed as described below.

2. METHODS

The purpose of this experiment was to assess the perceptual contribution of nasal murmur in different SNRs. To achieve this goal, the differences of perceptual effects *via* confusion patterns for each of the 16 nasal sounds were compared with and without nasal murmur being filtered out.

2.1. Subjects

Twenty college students with normal hearing participated in the study. All subjects were under 40 years old, and self-reported no history of speech or hearing disorder. By a short pilot test, it was found that they could perfectly



(A) (B)

Fig. (3). 3DDS analysis of the nasal consonants /m/ (**A**) and /n/ (**B**). (**a**) AI-gram with primary cues highlighted with red rectangles; (**b-d**) confusion patterns as a function of truncation time (**b**), SNRs (**c**), and cutoff frequency (**d**). (**e**) AI-grams of the consonant region, which is the region between solid vertical blue lines on panel (**a**) , under different SNRs. Nasal murmur is highlighted with red ellipsis.

recognize nasal consonants /m/ and /n/ in 12 dB SNR. The subjects were paid for their participation.

2.2. Speech Stimuli

The speech stimuli were eight /ma/ and eight /na/ syllables chosen from the University of Pennsylvania's Linguistic Data Consortium (LDC) LDC2005S22 "Articulation Index Corpus". Data from Phatak *et al.* [21] verified that these tokens had 0% recognition error at and above 12 dB SNR. To control the number of variables, the tokens selected were totally the same as used in previous three independent experiments of 3DDS [1].

2.3. Modification to Stimuli

To compare the differences between nasal consonants with and without nasal murmur, the nasal murmur region of 16 nasal consonants was filtered via a sixth order elliptical filter having stop band of 60 dB. The filters were implemented in MATLAB. Following this, different levels of white noise were added, so that all speech sounds were masked at 7 different SNRs: -18, -15, -12, -6, 0, 6, and 12 dB.

2.4. Experimental Procedure

Since the experimental set consisted of only two nasals, another 9 consonants into the presentations were added, so that the listeners would not deduce the experimental subset. These consonants are /p/, /b/, /f/, /d/, /k/, /s/, /v/, /g/, /z/, chosen from the University of Pennsylvania's Linguistic Data Consortium (LDC) LDC2005S22 "Articulation Index Corpus". Each of these consonants were spoken by six different talkers, and speech sounds were masked at 7 different SNRs: -18, -15, -12, -6, 0, 6, and 12 dB. So these sounds along with nasals brought 602 tokens (2 nasal consonants × 8 talkers × 7 SNRs × 2 conditions + 9 other consonants × 6 talkers × 7 SNRs = 602 tokens). All listening tests were performed in a single-walled, sound-proof testing booth (Acoustic System model number 27930) and the subjects listening to the speech sounds were at the most comfortable level using Sennheiser HD280 headphones. They were instructed to record the sound by selecting one of the 13 choices displayed in a GUI interface. These 13 choices included the two nasal consonants, the 9 other consonants, as well as two additional choices: "Noise Only" and "Others". This interface was programmed with MATLAB. All speech stimuli were presented to the listeners using a computer running the MATLAB procedure in Ubuntu 7.04 Linux system, with its CPU located outside of the testing booth to eliminate noise.

2.5. Practice Session

Before each experiment, a 3-min practice session was held using speech sounds with 18 dB SNR. The subjects would listen to the speech stimuli and make selection, then

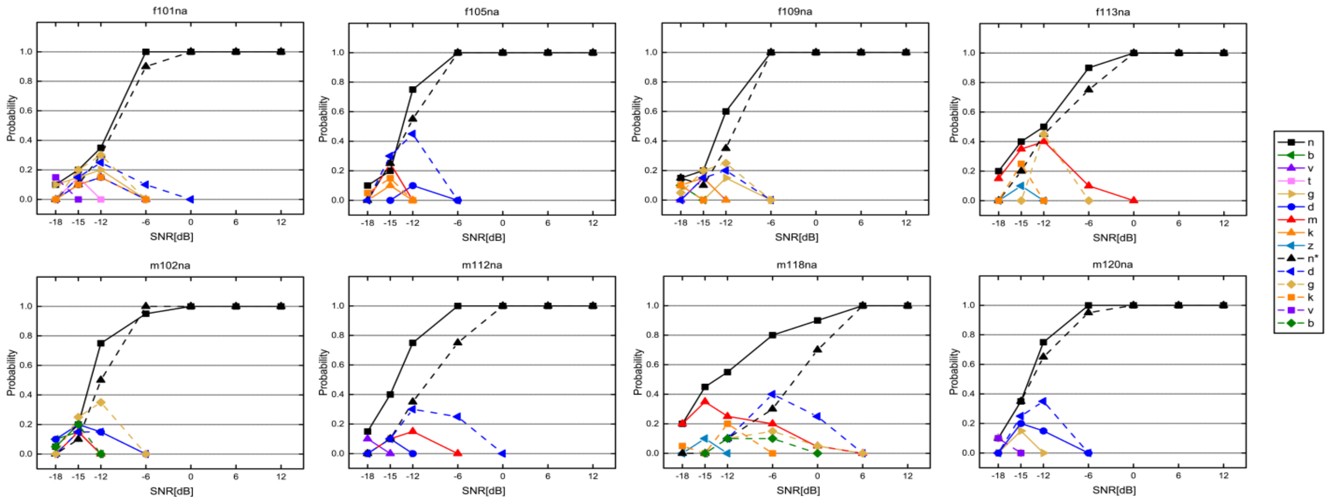


Fig. (4). Comparison of recognition scores of eight pairs of original and modified /m/ sounds. (/m*/ indicating the /m/ sound with nasal murmur filtered out)

the correct sound would be displayed on the screen. If their choice was right, it would move on to the next sound and this sound would not be played again. But if their choice is wrong, this sound would be inserted back to the rest of the playlist at a random position, so that it can be heard again later, until the subject makes the right choice on its turn.

2.6. Experimental Session

For the experimental session, every one of these 602 tokens was played randomly once. If the subject didn't hear a particular token clearly and wanted to hear it again, they could use the "repeat" bottom for a maximum of 3 listening times. Tokens were presented every 5 seconds, and for every 15 min of experiments there was a 5 to 10 min break to avoid fatigue. Subjects could also pause when they needed an additional break. Different from the practice session, the experimental session did not have any feedback on the correctness of their choices. The response was recorded automatically into the experiment database.

3. RESULTS AND ANALYSIS

The data was first plotted in the form of confusion patterns, which are defined as the proportion of all responses for a particular token as a function of SNR [22]. Each entry of the table represents the proportion of the sound on the left-most column and is reported to be heard among all responses under a specific SNR, as indicated on the top of each column. As shown in Table **1**, when a sound was played m115/ma/ under different SNRs, the subjects reported it was /ma/, /pa/ or /na/. The corresponding percentage of responses under different SNRs is shown in every column. The sum of some columns is not equal to 1 because "Noisy only" and "others" options are not shown in the table.

Moreover, for easier observation, each confusion matrix is plotted as line plots. Fig. (**4**) represents the line plots for all the eight /m/ tokens, and Fig. (**5**) indicates the line plots for all the eight /n/ tokens. In each line plot, the solid line represents the percentage of a specific consonant heard and reported when the original nasal sound was played under different SNRs; each dashed line represents the percentage of a specific consonant heard and reported when the nasal sound with its nasal murmur filtered out was played under different SNRs.

Table 1. Confusion matrix for m115ma.

m115ma	SNR [dB]						
	-18	-15	-12	-6	0	6	12
m	0.1	0.2	0.4	0.95	1	1	1
p	0	0	0.2	0	0	0	0
n	0	0.3	0	0	0	0	0

It can be seen from Figs. (**4** and **5**) that when SNRs are at high levels, both the recognition score of nasal consonant with murmur (represented by black solid lines) and the recognition score of nasal consonant without nasal murmur (represented by black dash lines) are equal to 100%. It means that when the primary cue of nasal consonant is heard clearly, listeners can correctly recognize nasal consonants even with their nasal murmur filtered out. This phenomenon is consistent across different talkers.

It can also be seen from Figs. (**4** and **5**) that when the recognition score of nasal consonant with murmur started to drop, the recognition score of nasal consonant without nasal murmur dropped much quicker than the original nasal [Fig. (**4**)]. Comparison of recognition scores of eight pairs of original and modified /m/ sounds (/m*/ indicating the /m/ sound with nasal murmur filtered out) with nasal murmur is also visible in Fig. (**6**), showing the average percentage decrease of recognition scores across the eight tokens

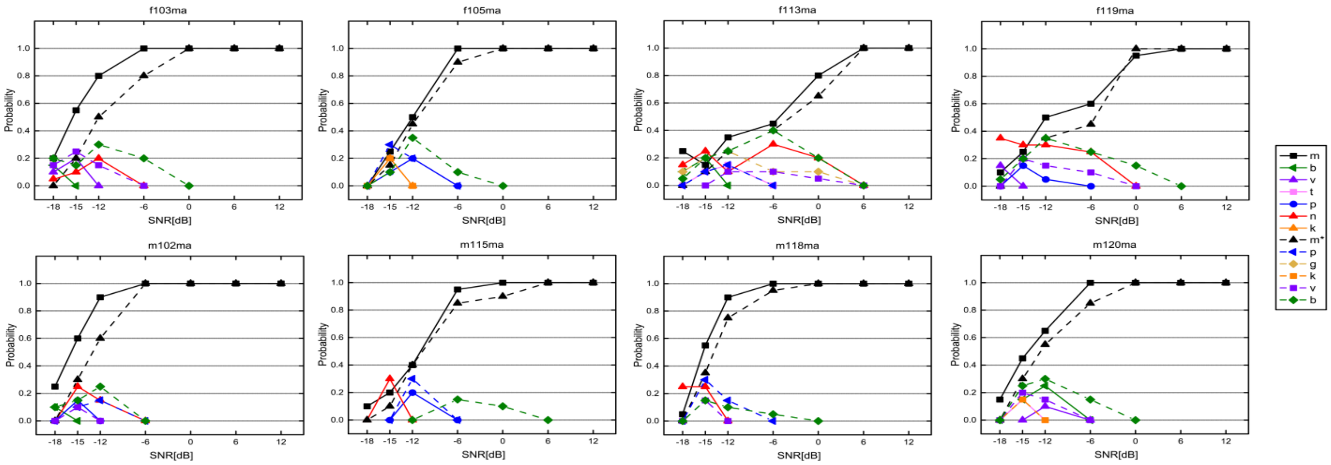


Fig. (5). Comparison of recognition scores of eight pairs of original and modified /n/ sounds. (/n*/ indicating the /m/ sound with nasal murmur filtered out).

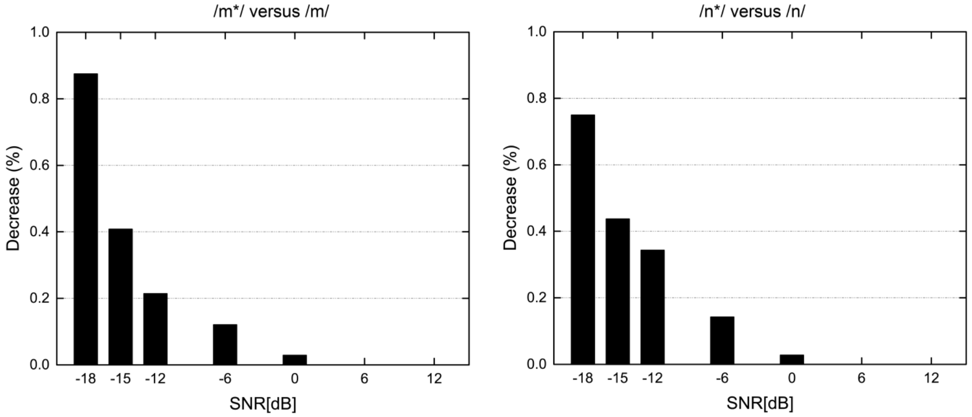


Fig. (6). Percentage decrease in recognition scores for /m/ and /n/ when nasal murmur is filtered out. (*indicating the modified nasal with its nasal murmur filtered out).

at different SNRs. It is also evident that at the high SNR end, the recognition score of nasals with murmur are not better than nasals without murmur, since they are all equal to 100%. While at the low SNR end, compared with the recognition score of original nasals, the recognition score of nasals without murmur has a significant decrease. It means that when the primary cue of nasal consonant is not heard clearly, listeners can get perceptual support from nasal murmur.

Such support becomes more effective and useful as the SNR decreases, in accordance with the fact that the difference in recognition scores becomes larger as the SNR becomes lower. This is because nasal murmur usually has much stronger energy than primary cue. When the primary cue is gradually masked by white noise, the nasal murmur can still be heard clearly. As a result, its contribution to the correct perception of nasal consonant gradually increases.

We can take m118/ma/ as an example. In Fig. (**3A**), when SNRs are 0, 6, 12 dB, the primary cue in red rectangle is very clear in AI-grams; and in Fig. (**4**), in the m118 plot, there is no difference in recognition scores between the nasal with murmur (m) and the nasal without murmur (m*). The recognition scores are both equal to 100%. People do not need to hear nasal murmur for correct perception. On the other hand, in Fig. (**3A**), when SNRs are -6, -12, -15 dB, the murmur is still clear but the primary cue becomes more washed out in the corresponding AI-gram; and in the m118 plot of Fig. (**4**), the difference in recognition scores between the nasals with murmur (m) and the nasal without murmur (m*) becomes bigger and bigger. It indicates that under these SNRs, the nasal murmur gives a supplement to the primary cue.

CONCLUSION

Although it is not easy to identify the precise locations of perceptual cues, a number of techniques have been applied to locate these cues in our previous research. Especially by integration of three different independent measures and the

AI-gram as the 3DDS method, the primary perceptual cue of nasal consonants: /m/ and /n/ are also identified. In this research, our previous result is extended based on nasal consonants. After further experimentation and analysis, it has been found that the contribution of nasal murmur to correct perception of nasal consonants depends on different situations. In general, under high SNRs when the primary cue can be heard clearly, listeners can correctly identify the nasal consonants without nasal murmur. But under low SNRs, when the primary cue cannot be heard clearly, nasal murmur then effectively gives supplement to the primary cue for correct perception of nasal consonants /m/ and /n/, and can greatly reduce the confusions.

CONFLICT OF INTEREST

The author confirms that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

The authors would like to thank Dr. Jont Allen,Dr. Feipeng Li, Dr. Andrea Trevino, and other members of the Human Speech Research group at the University of Illinois for their help and discussions. This research was partially supported by a University of Illinois grant.

REFERENCES

[1] F. Li, A. Menon, and J.B. Allen, “A psychoacoustic method to find the perceptual cues of stop consonants in natural speech”, *The Journal of the Acoustical Society of America,* vol. 127, no. 4, pp. 2599-2610, 2010.

[2] F. Li, A. Trevino, A. Menon, and J.B. Allen, “A psychoacoustic method for studying the necessary and sufficient perceptual cues of American English fricative consonants in noise”, *The Journal of the Acoustical Society of America,* vol. 132, no. 4, pp. 2663-2675, 2012.

[3] F. Li, “Perceptual cues of consonant sounds and impact of sensorineural hearing loss on speech perception”, Urbana, IL. University of Illinois at Urbana-Champaign, 2009.

[4] F. Li, and J.B. Allen, “Multiband product rule and consonant identification”, *The Journal of the Acoustical Society of America,* vol. 126, no. 1, pp. 347-353, 2009.

[5] R.A. Cole, and B. Scott, “Towards a theory of speech perception”, *Psychological Review,* vol. 81, pp. 348-374, 1974.

[6] F. Cooper, C. D. Pierre, M. L. Alvin, M. B. John, and J. G. Louis, “Some experiments on the perception of synthetic speech sounds”, *The Journal of the Acoustical Society of America*, vol. 24, pp. 579-606, 1952.

[7] R.K. Potter, G.A. Kopp, and H.G. Kopp, *Visible Speech*. New York: Dover Publication, 1966.

[8] G.W. Hughes, and M. Halle, “Spectral properties of fricative consonants”, *The Journal of the Acoustical Society of America*, vol. 28, no. 2, pp. 303-310, 1956.

[9] A. Liberman, “Some results of research on speech perception”, *The Journal of the Acoustical Society of America*, vol. 29, pp. 117-123, 1957.

[10] J. Heinz, and K. Stevens, “On the properties of voiceless fricative consonants”, *The Journal of the Acoustical Society of America*, vol. 33, no. 5, pp. 589-596, 1961.

[11] S.E. Blumstein, K.N. Stevens, and G.N. Nigro, “Property detectors for bursts and transitions in speech perceptions”, *The Journal of the Acoustical Society of America*, vol. 61, no. 5, pp. 1301-1313, 1977.

[12] K.N. Stevens, and S.E. Blumstein, “Invariant cues for place of articulation in stop consonants”, *The Journal of the Acoustical Society of America*, vol. 64, no. 5, pp. 1358-1368, 1978.

[13] D. Recasens, “Place cues for nasal consonants with special reference to Catalan”, *The Journal of the Acoustical Society of America*, vol. 73, pp. 1346-1353, 1983.

[14] R.V. Shannon, F. G. Zeng, V. Kamath, J. Wygonski, and M. Ekelid, “Speech recognition with primarily temporal cues”, *Science,* vol. 270, no. 5234, pp. 303-304, 1995.

[15] P. Delattre, A. Liberman, and F. Cooper, “Acoustic loci and translational cues for consonants”, *The Journal of the Acoustical Society of America*, vol. 24, no. 4, pp. 769-773, 1955.

[16] R. Remez, P.E. Rubin, D.B. Pisoni, and T.D. Carrell, “Speech perception without traditional speech cues”, *Science,* vol. 212, no. 4497, pp. 947-949, 1981.

[17] V. Hazan, and S. Rosen, “Individual variability in the perception of cues to place contrasts in initial stops”, *Perception & Psychophysics*, vol. 59, no. 2, pp. 187-200, 1991.

[18] M.S. Régnier, and J.B. Allen, “A method to identify noise-robust perceptual features: Application for consonant”, *The Journal of the Acoustical Society of America*, vol. 123, pp. 2801-2814, 2008.

[19] B.E. Lobdell, *Models of Human Phone Transcription in Noise Based on Intelligibility Predictors*, Urbana, IL. University of Illinois at Urbana-Champaign, 2009.

[20] B. E. Lobdell, J. B. Allen, and M.A. “Hasegawa-Johnson, Intelligibility predictors and neural representation of speech”, *Speech Communication*, vol. 53, no. 2, pp. 185-194, 2011.

[21] S.A. Phatak, A. Lovitt, and J.B. Allen, “Consonant confusions in white noise”, *The Journal of the Acoustical Society of America,* vol. 124, no. 2, pp. 1220-33, 2008.

[22] J.B. Allen, “Consonant recognition and the articulation index”, *The Journal of the Acoustical Society of America*, vol. 117, no. 4, pp. 2212-2223, 2005.

Received: February 18, 2015 Revised: March 22, 2015 Accepted: March 31, 2015